ENVIRONMETRICS
Environmetrics 2004; **15**: 119–128 (DOI: 10.1002/env.627)

Optimization of ordered distance sampling‡

Ryan M. Nielson[1,2,3], Robert T. Sugihara[4], Thomas J. Boardman[2] and Richard M. Engeman[1]*,†

[1]*National Wildlife Research Center, USDA/Animal and Plant Health Inspection Service/Wildlife Services, 4101 Laporte Avenue, Fort Collins, CO 80521-2154, U.S.A.*
[2]*Colorado State University, Statistics Department, Fort Collins, CO 80523-1877, U.S.A.*
[3]*Western Ecosystems Technology, Inc., 2003 Central Avenue, Cheyenne, WY 820001, U.S.A.*
[4]*National Wildlife Research Center, USDA/Animal and Plant Health Inspection Service/Wildlife Services, Hawaii Field Station, P.O. Box 10880, Hilo, HI 96721, U.S.A.*

SUMMARY

Ordered distance sampling is a point-to-object sampling method that can be labor-efficient for demanding field situations. An extensive simulation study was conducted to find the optimum number, g, of population members to be encountered from each random starting point in ordered distance sampling. Monte Carlo simulations covered 64 combinations of four spatial patterns, four densities and four sample sizes. Values of g from 1 to 10 were considered for each case. Relative root mean squared error (RRMSE) and relative bias were calculated for each level of g, with RRMSE used as the primary assessment criterion for finding the optimum level of g. A non-parametric confidence interval was derived for the density estimate, and this was included in the simulations to gauge its performance.

Superior estimation properties were found for $g > 3$, but diminishing returns, relative to the potential for increased effort in the field, were found for $g > 5$. The simulations showed noticeable diminishing returns for more than 20 sampled points. The non-parametric confidence interval performed well for populations with random, aggregate or double-clumped spatial patterns, but rarely came close to target coverage for populations that were regularly distributed. The non-parametric confidence interval presented here is recommended for general use. Published in 2004 by John Wiley & Sons, Ltd.

KEY WORDS: density; distance sampling; point-to-object sampling; spatial pattern

1. INTRODUCTION

In many areas of ecology it is necessary to estimate the density of stationary objects in the field such as plant communities, points of crop damage, bird nests or animal burrows. An ideal density sampling method would produce an unbiased estimate, be robust to different population spatial patterns and densities, and be easily applied in various field situations. Two general sampling techniques that produce density estimates are the quadrat method and the various types of distance, or plotless, methods. The well-known quadrat method involves randomly locating plots, or quadrats, of a given

*Correspondence to: Dr Richard M. Engeman, National Wildlife Center, 4101 Laporte Avenue, Fort Collins, Co 80521-2154, U.S.A.
†E-mail: Richard.M.Engeman@usda.gov

Published online 27 January 2004
Published in 2004 by John Wiley & Sons, Ltd.
Received 20 March 2002
Accepted 28 April 2003

size and thoroughly searching for every population member in each quadrat. This technique produces unbiased estimates and is robust over population spatial patterns, given an appropriate quadrat size and that all individuals within each quadrat are counted; each object within each quadrat must be detected with probability one.

Examples of distance methods include the variable area transect, angle-order and the very popular line transect method (Burnham *et al.*, 1980; Buckland *et al.*, 1993; Krebs, 1998). Although line transect sampling has proven to be a method that provides reliable estimates of density under fairly mild assumptions (Buckland *et al.*, 1993), there are a few field situations that present difficulties for the researcher such as when the objects are very dense or the landscape is very dense and it is difficult to follow an assigned path and assume that all objects on the line transect are detected with probability one. Quadrat sampling can also become arduous when it is difficult to make an accurate count of all members within a quadrat, such as sampling heavily damaged crops.

Cottam (1947) introduced distance sampling as a method for more easily obtaining density estimates under these circumstances. Recently Engeman *et al.* (1994) made comparisons of the performances of 25 distance methods on different spatial patterns, densities and sample sizes. Recommendations were made as to which PDEs provided the best estimation properties while remaining feasible to apply in difficult field situations. This work follows up on that paper by optimizing one of the better performing, but simplest to apply, of the methods from that study—the ordered distance (OD) method originally derived by Morisita (1957) and further developed by Pollard (1971).

Buckland *et al.* (1993) recommends against using point-to-object methods, 'except in special cases, such as estimating the density of forest stands.' The current article investigates the quality of OD density estimates not only because ordered distance sampling is used in forestry, but also because it holds potential for very broad application. For example, it is receiving attention by some agricultural researchers who need to quickly estimate the relative density of points of crop damage in situations where it would be too labor-intensive or otherwise unrealistic to use quadrat or line-transect methods. While this investigation is motivated by situations where quadrat or line transect sampling may prove too tenuous or labor-inefficient for the desired objective, the results are applicable in general.

2. ORDERED DISTANCE

Ordered distance sampling is performed by first randomly locating a sample size of n points in the area of interest. At each of these n locations the researcher searches for the gth nearest population individual from the random starting point and then records the exact distance from the point to the gth individual. This could be applied by searching an ever-widening radius out from the random starting point, or by simply visually identifying what appears to be the gth closest individual, and then verifying this by distance measurement. If we let g be the number of nearest population individuals searched for from each of the n random starting points, and $R_{(g)i}$ be the distance to the gth individual from the ith random starting point, then the formula for the OD estimator given by Pollard (1971) is

$$\hat{D} = (ng - 1) \Big/ \left[\pi \sum (R_{(g)i})^2\right]$$

This estimator was derived for sampling from random spatial patterns and is unbiased for estimates from such populations. The variance estimate for this estimator is

$$\widehat{\text{var}}(\hat{D}) = (\hat{D})^2/(ng - 2)$$

Two major assumptions need to be considered when using this estimator: (i) the population under investigation follows a random spatial distribution and (ii) the g nearest population members to each random point must be detected with probability one. This study seeks to understand how the estimator performs under spatial patterns other than the random, but it does not address detectability, assumption (ii). In previous simulation studies, Engeman *et al.* (1994) considered using $g = 1$, 2 or 3. Presented here is an investigation into the estimation properties when sampling populations that represent a diversity of spatial patterns and at a variety of densities. In each circumstance, estimation based on locating $g = 1, 2, \ldots, 10$ population members from each random starting point is considered. Then, attempts are made to determine optimal values of g and n, or values that represent points of diminishing return for the field researcher.

3. SIMULATION STUDY DESIGN

A simulation program was written in Fortran 77 (Version 5.0, MS-DOS operating system), each run of which was specified by a combination of population spatial pattern, population density and sample size (number of random OD starting points). Under examination were 64 combinations encompassing four spatial patterns, four densities and four sample sizes. At each random sampling point, the distances to the 10 nearest individuals were measured. Thus, estimates could be calculated using $g = 1$, 2, 3, 4, 5, 6, 7, 8, 9 and 10. This results in a simulation study on four features that potentially could affect estimation by ordered distance sampling: spatial pattern, density, sample size, and number of population members searched for at each random sampling point.

The uniform random number generator used for placing population individuals and locating sampling points was the UNIF routine (Bratley *et al.*, 1983). Where required, the VNORM routine (Bratley *et al.*, 1983) was used to convert the uniform random numbers to normal random numbers. UNIF has been extensively tested for uniformity, independence and non-periodicity of the numbers generated and VNORM tested for accuracy (Brody and Morais, 1987).

The density used in a particular run of the program was specified by inputting the size of a rectangular area (the length of each dimension) and the number of individuals to reside in that area. Target population densities of 2, 5, 10 and 20 individuals per unit area were examined. The area used for each density was large enough to ensure that the target population was several orders of magnitude larger than the number of sampling points.

Four spatial patterns were considered for the populations simulated in this study: random, regular, aggregate and double-clumped. The *random* pattern (also called Poisson, in recognition that the points are distributed as a two-dimensional Poisson process) was simulated by generating the appropriate number of random co-ordinates from a uniform distribution in the designated area. The *regular* spatial pattern was generated by dividing the area into a grid of rectangles, the same number as individuals in the population. The population members were then situated by randomly locating one individual in each rectangle. For the *aggregate* pattern, the centers of user-specified number of clumps were randomly located in the designated area. In addition to the clump center point, a user-specified number of 'offspring' for the clumps were located within the user-specified radius of the center (parent) point. These offspring were located within the clump about the parent point using co-ordinates randomly generated from the standard bivariate normal distribution. This tended to concentrate the members of the clump near the center point. Aggregate spatial patterns approximate many naturally occurring biological population patterns. For the simulations, clumps were comprised of five individuals (the center point or 'parent' and four 'offspring') that were severely clumped, with the offspring located

within a clump radius of 15 distance units. The pattern defined as *double-clumped* is a second-order aggregation that was generated in a similar fashion to the aggregate pattern. The difference is that, for the double-clumped pattern, the individuals in the clumps of the aggregate pattern are used for center points (parents) for sub-clumps of two individuals. The two individuals of the sub-clumps include the parent plus one other point (offspring) randomly generated from the standard bivariate normal distribution. The radius for the sub-clump is restricted to one-half that for the clump (7.5 units). This spatial pattern approximates some of the field patterns that we have observed for rodent burrows and animal damage locations. It also provides one of the severest tests of the estimation method.

Randomly located starting points for the OD samples were required to initiate the sampling procedures. The sample sizes considered in this study refer to the number of such random starting points placed in the population. Sample sizes of random starting points examined were 5, 10, 20 and 40.

There were two runs of the simulation program for each spatial pattern-by-density-by-sample-size combination. At each replication of each run of the program, a new population was generated and a new set of random sampling points applied. Each simulation run comprised 5000 such replications. The observed statistics accumulated over the 5000 replications at each g included the mean density estimate, variance, relative bias, mean squared error (MSE) and relative root mean squared error (RRMSE). RRMSE was calculated as

$$RRMSE = \left\{ \left[\sum (\hat{D} - D)^2 \Big/ D^2 \right] \Big/ I \right\}^{1/2}$$

where $\hat{D}$ was the estimated density, D was the true density, and $I = 5000$ was the number of replications in the simulation run. RRMSE was used as the primary criterion for comparing the performance of the estimates (see, for example, Patil *et al.*, 1979; Engeman and Bromaghin, 1990; Engeman *et al.*, 1994; Engeman and Sugihara, 1998), because it encompasses variance and bias, and it is unitless. Also calculated was the relative bias (RBIAS)—the mean observed bias divided by the true parameter value. The statistics presented from these simulations are 'relative' statistics (divided by the true density) to standardize the scale across the density parameter being estimated.

The trade-off between n and g was examined by developing a predictive model of RRMSE for the density estimate. Only variables under the investigator's control (n and g) were used to develop predictive equations. Density, therefore, was not included. Also, it was presumed that: (i) different relationships between RRMSE and n and g might come into play for different spatial patterns of population members, and (ii) a field investigator might have information on the population pattern, based on similar populations elsewhere or on other a priori knowledge. For each spatial pattern, the equations with the lowest value of Akaike's Information Criterion (AIC) (Akaike, 1973; Burnham and Anderson, 1998) were chosen, looking at all reasonable subsets of n, g, n^2 and g^2, and the product ng. No prior knowledge of spatial patterns was also considered, and thus all RRMSE results were analyzed together.

Confidence interval coverage based on two calculation methods for producing $1-\alpha$ confidence intervals were also examined for each spatial pattern-by-density-by-sample-size combination. The first formula used a normal approximation:

$$\hat{D} - N_{1-\alpha/2}\sqrt{\widehat{\text{var}}(\hat{D})} \leq D \leq \hat{D} + N_{1-\alpha/2}\sqrt{\widehat{\text{var}}(\hat{D})}$$

The second confidence interval formula was a non-parametric method based on applying a confidence interval to the median (e.g. Hollander and Wolfe, 1973) from a sample of $nR_{(g)i}$s. The calculation steps for a $1-\alpha$ confidence interval on density from ordered distance sampling are given in the Appendix.

For each of the calculation methods, 90 and 95% confidence intervals were considered. Confidence interval coverage for a given target coverage, n, and g, was calculated as the percentage of the 5000 iterations for which confidence intervals contained the true population density.

4. RESULTS AND DISCUSSION

The results from the 128 population simulations (two runs each of the 64 combinations) are summarized as the mean RRMSE in each spatial pattern at each sample size in Figure 1 and the mean RBIAS in Figure 2. Besides the extensive testing involved in developing a simulation program, various aspects of the results confirm that the program was functioning properly. Each of these 64 combinations was run twice, with different input seeds for the random number generator. There was very little variability between simulation runs, with some of the resulting statistics often being identical (to four decimal places) between the two runs. The mean relative bias (RBIAS) for the random spatial pattern is zero for all densities, sample sizes and number of nearest individuals, g, which also confirms that the simulations were running properly (Pollard, 1971).

For most of the simulation combinations, definite improvement in estimation was made with $g > 3$, although the relative magnitudes of the improvements usually diminished after $g = 5$ (Figures 1 and 2).

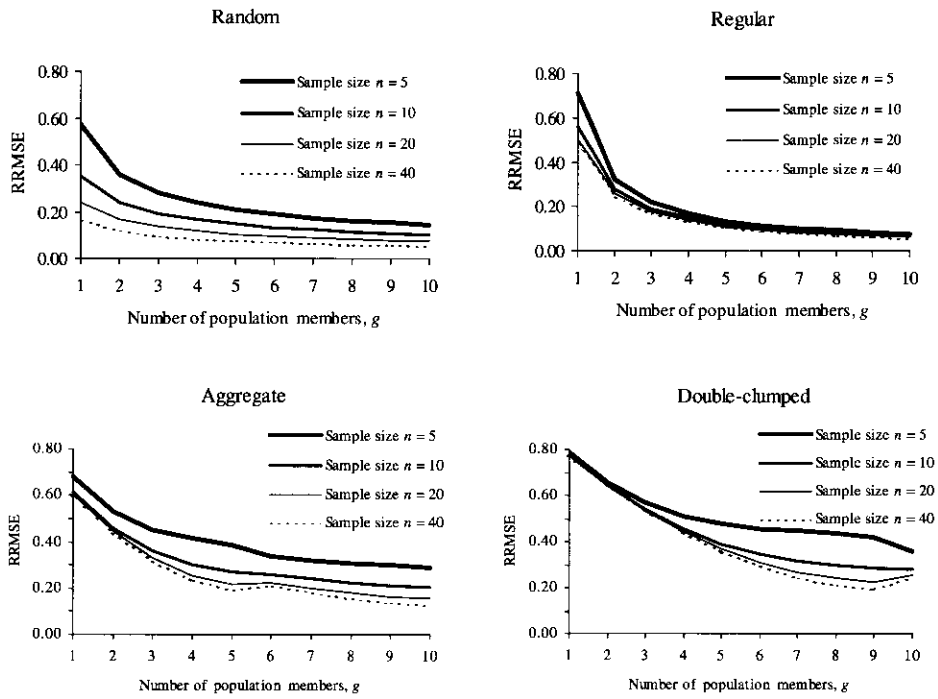

Figure 1. Mean RRMSE results for each spatial pattern and sample size using $g = 1$ to 10, across all densities

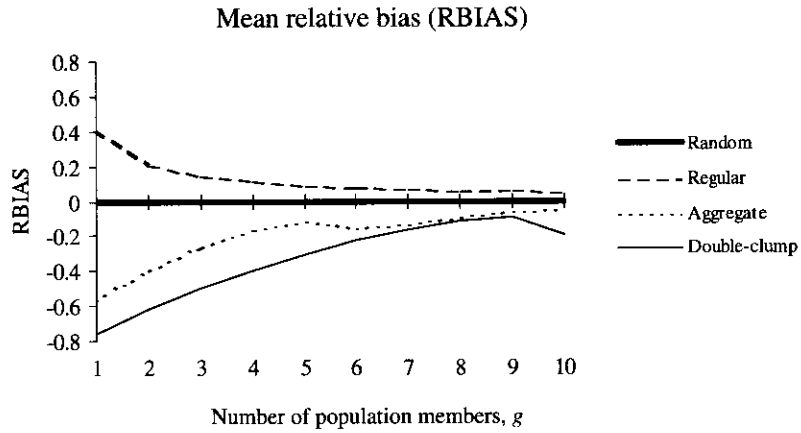

Figure 2. Mean relative bias (RBIAS) for each spatial pattern using $g = 1$ to 10, across all densities and sample sizes

Across all spatial patterns, densities and sample sizes, the RRMSE was reduced by about 60% by measuring the distance to the 5th nearest individual, compared to using $g = 1$ (Table 1). A further reduction of only 12% is obtained by measuring the distance to the 10th nearest individual.

As would be expected, estimation improved with increasing sample size. RRMSEs did not substantially change with density. Figure 1 shows that the RRMSE was much lower for the random and regular patterns than for the populations where individuals were aggregated.

RRMSE predictive equations are shown in Table 2. For the random, regular and aggregated spatial patterns the best equations were produced by full second order models with an interaction term (n, g, n^2, g^2 and ng). Second order polynomials without interaction terms were found to be the best fitting models for the double-clump pattern and the 'overall' model.

Mean RBIAS was zero for the random spatial pattern, positive for the regular pattern and negative for the aggregate and double-clumped patterns. The results for the aggregate and double-clump spatial patterns could possibly be explained by the size of the clumps, or clusters, in the populations (cluster size equaled five for the aggregate and nine for the double-clumped pattern).

Estimates of density based on sample sizes of 5 and 10 were not very reliable (large RRMSE) compared to estimates based on larger samples sizes of 20 and 40 (Figures 1 and 2), and a sample size

Table 1. Percentage reduction of RRMSE in comparison to using $g = 1$, across all spatial patterns, densities and sample sizes

g	RRMSE	Reduction (%)
1	0.57	0
2	0.40	30
3	0.32	44
4	0.27	53
5	0.23	60
6	0.21	63
7	0.19	67
8	0.17	70
9	0.16	72
10	0.16	72

Table 2. Predictive equations for RRMSE based on sample sizes and levels of g

Pattern	Parameter estimates					
	Intercept	Sample size	g	(Sample size)^2	g^2	(Sample size) $\times$ g
Random	0.57687	−0.01747	−0.07895	0.00021	0.00417	0.00064
Regular	0.69815	−0.00700	−0.15640	0.00008	0.00996	0.00036
Aggregate	0.81405	−0.01270	−0.11687	0.00023	0.00734	−0.00021
Double-clump	0.75448	−0.01114	−0.11859	0.00018	0.00728	—
Overall	0.75571	−0.01117	−0.11904	0.00018	0.00731	—

of $n = 20$ seemed to be the point of diminishing returns for increased effort in the field. Thus, confidence interval coverage by two methods is examined only for sample sizes of 20, and results are displayed in Figures 3 and 4. The normal confidence interval produced excellent coverage in the random spatial patterns for all values of g from 2 to 10. This method provided inconsistent results among target coverages for the regular spatial pattern. The results for $n = 20$ and $g > 2$ were much higher than target coverage (very conservative). The normal confidence interval method produced extremely poor results for both aggregate and double-clump spatial patterns for all values of g.

The non-parametric confidence interval produced good coverage for the random spatial pattern. It should be noted that coverage for $n = 40$ was significantly lower than target coverage for values of $g < 5$ (better coverage for $g > 4$). The regular spatial pattern was a challenge for this method, as

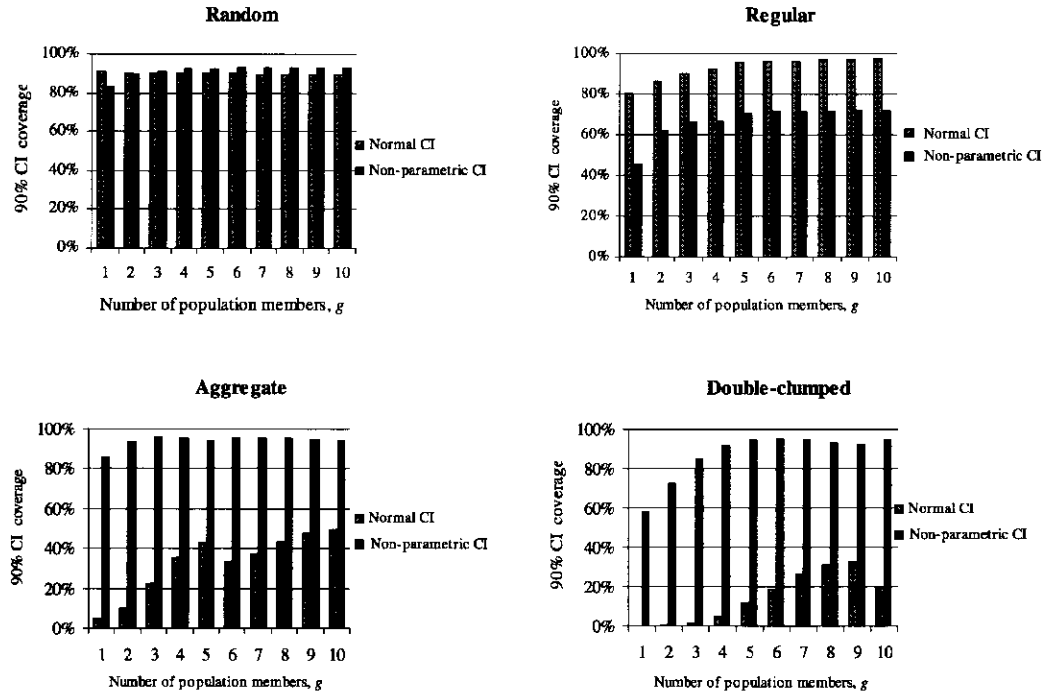

Figure 3. 90% confidence interval coverage for each spatial pattern using a sample size of 20 and $g = 1$ to 10, across all densities

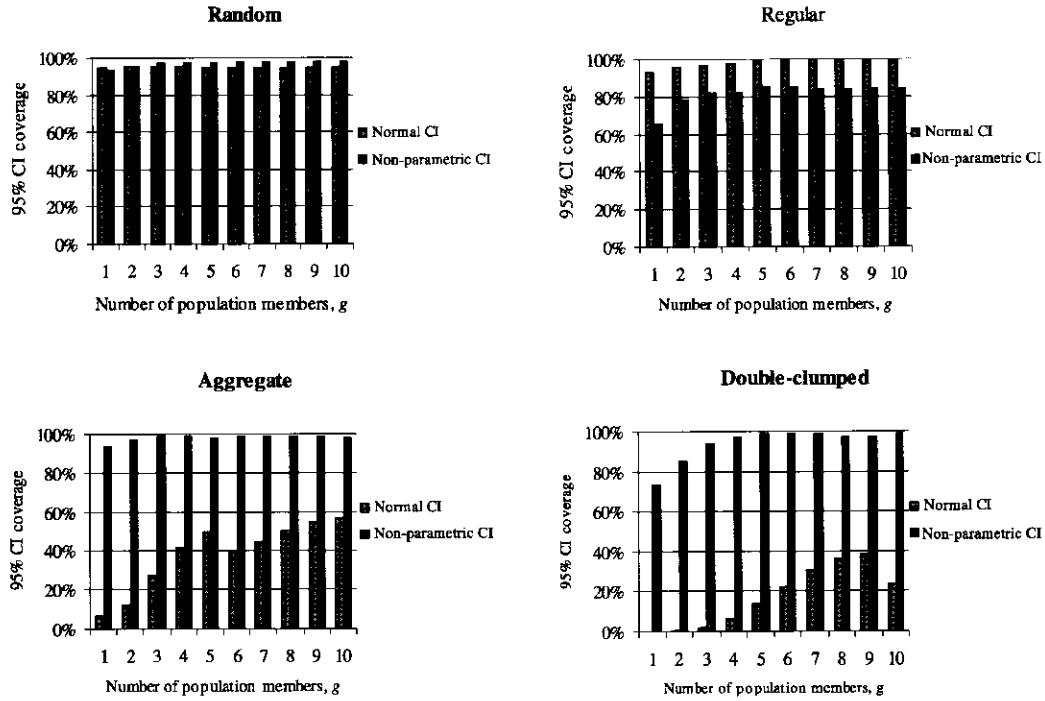

Figure 4. 95% confidence interval coverage for each spatial pattern using a sample size of 20 and $g = 1$ to 10, across all densities

coverage results were always short of target. In contrast, the non-parametric confidence interval method performed much better than the normal method for both the aggregate and double-clump spatial patterns, although its coverage was much higher than target for both patterns when g was >2 or 3.

5. CONCLUSIONS

Although using $g \leq 3$, as originally proposed by Pollard (1971), does produce good density estimates for the random spatial pattern, in most situations the field researcher would benefit substantially from locating the 5th nearest population individual from each random starting point. This becomes more crucial as the population spatial pattern deviates more from the random pattern. It is not clear how reasonable using $g = 5$ is in all field situations. This is where the predictive equations for the RRMSE in Table 2 may prove valuable. The researcher could weigh the costs and benefits of using different combinations of g and n, knowing the specific field situation, and possibly even the spatial pattern of the population under investigation. When considering a sample size to use for OD sampling, it should be noted that $n = 20$ seems to be the point of diminishing returns for lowering the RRMSE and RBIAS of the estimator. We suggest first considering a level of $n \simeq 20$ for any field investigation using the OD method, and then increasing g rather than n if time and costs permit. However, if field conditions are 'easy' enough to use $g > 6$ or $n > 20$, then quadrat or line transect sampling should be considered.

Confidence interval coverage is very close to target for the non-parametric confidence interval when using $g = 5$ and $n = 20$, for the various spatial patterns and densities. If the investigator has

information as to which spatial pattern the population tends to follow, then a choice can be made between the normal confidence interval and the non-parametric confidence interval methods. If it is known that the population follows a random or regular spatial pattern, then the normal confidence interval would provide coverage closest to target. The regular spatial pattern is not a very common occurrence in nature, but it may be encountered when investigating populations of territorial individuals over a homogenous landscape, such as colonial nesting waterbirds. If the population follows the aggregate or double-clumped patterns, which are more common in field populations, such as for plant communities, crop damage and animal burrows, then the non-parametric confidence interval should be used. It might be possible for the investigator to first use the ordered distance method to sample from the population, and then calculate Pielou's index of non-randomness (Pielou, 1959) to indicate which spatial pattern is being observed—regular, random or aggregate—and then choose a confidence interval method.

It should be noted that non-parametric confidence intervals tend to be relatively large (wide) for a given target coverage, and when sampling from lower density populations, the confidence interval can encompass zero. This can be easily remedied, without losing accuracy, by replacing the lower endpoint of the non-parametric confidence interval with the observed density. Other suggestions for improved confidence interval methods to investigate include using a bootstrap variance and confidence interval if sample sizes are adequate (say $n \geq 20$), or derivation of a new confidence interval that is a combination of the normal and the non-parametric confidence interval methods (Engeman and Sugihara, 1998).

We sought to optimize the application of OD sampling through extensive simulations. In natural situations the pattern and density of populations can vary greatly and a simulation using artificial populations can only approximate natural processes. Therefore, the true test of the results presented here would be verification using a wide variety of fully enumerated field data sets. At present, it appears that OD sampling using $g = 4$ or $g = 5$ offers the field investigator a method that is relatively easy to apply, confusion free, and produces quality estimates for populations exhibiting the random, regular or aggregated spatial patterns. Although ordered distance may be the favored method when labor-efficiency is preferred over unbiasedness or in extreme field situations such as estimating the intensity of crop damage, it is important to remember the limitations of the estimator and its assumptions. Given the information in this paper on the size and direction of bias and estimator efficiency, the field researcher can make informed decisions regarding the use of the ordered distance method and the resulting density estimate for the investigated population.

APPENDIX

The calculation method for a non-parametric $(1-\alpha) \times 100\%$ confidence interval on density from ordered distance sampling, when the distance to the gth nearest population member from each random sampling point is measured, is as follows (at any given random sampling point let $R_{(g)i}$ = the distance from the ith sampling point to the gth population member; let n be the number of such random sampling points from which ordered distances were applied):

1. Order the n R_i from smallest to largest such that $R_i^{(1)} < R_i^{(?)} < \cdots < R_i^{(n)}$.
2. Calculate the value of C to the nearest integer, where

$$C = (n/2) - z_{\alpha/2}(n/4)^{1/2}$$

3. Let $\theta_1 = \left(R_i^{(C)}\right)^2$ and $\theta_2 = \left(R_i^{(n+1-C)}\right)^2$.
4. The lower (LL) and upper (UL) $1 - \alpha$ confidence limits are, respectively:

$$LL = (ng - 1)/(n\pi\theta_2)$$

$$UL = (ng - 1)/(n\pi\theta_1)$$

REFERENCES

Akaike H. 1973. Information theory as an extension of the maximum likelihood principle. In *Second International Symposium on Information Theory*, Petrov BN, Csaki F (eds). Akademiai Kiado: Budapest; 267–281.

Bratley P, Fox BL, Schrage LE. 1983. *A Guide to Simulation*. Springer-Verlag: New York, NY, USA.

Brody M, Morais D. 1987. UNIF is fast, reliable random number generator for IBM-AT microcomputer. *Research Information Bulletin Number 87–84*. USDI Fish and Wildlife Service, Washington, DC, USA.

Buckland ST, Anderson DR, Burnham KP, Laake JL. 1993. *Distance Sampling*. Chapman & Hall: London, UK.

Burnham KP, Anderson DR. 1998. *Model Selection and Inference*. Springer: New York, NY, USA.

Cottam G. 1947. A point method for making rapid surveys of woodlands. *Bulletin of the Ecological Society of America* **28**: 60.

Burnham KP, Anderson DR, Laake JL. 1980. Estimation of density from line transect sampling of biological populations. *Wildlife Monographs* **72**: 1–202.

Engeman RM, Bromaghin JF. 1990. An approach to estimating density from line transect data where animals move in response to the observer. *Journal of Statistical Computation and Simulation* **34**: 145–154.

Engeman RM, Sugihara RT. 1998. Optimization of variable area transect sampling using Monte Carlo simulation. *Ecology* **79**: 1425–1434.

Engeman RM, Sugihara RT, Pank LF, Dusenberry WE. 1994. A comparison of plotless density estimators using Monte Carlo simulation. *Ecology* **75**: 1769–1779.

Hollander M, Wolfe DA. 1973. *Nonparametric Statistical Methods*. Wiley: New York, NY, USA.

Krebs CJ. 1998. Ecological Methodology. Addison Wesley Longman: Menlo Park, CA, USA.

Morisita M. 1957. A new method for the estimation of density by spacing method applicable to nonrandomly distributed populations. (In Japanese: available as Forest Service translation Number 11116, USDA Forest Service, Washington, DC, USA.) *Physiology and Ecology* **7**: 134–144.

Patil SA, Burnham KP, Kovner JL. 1979. Nonparametric estimation of plant density by the distance method. *Biometrics* **35**: 597–604.

Pielou EC. 1959. The use of point-to-plant distances in the study of the pattern of plant populations. *Journal of Ecology* **47**: 607–613.

Pollard JH. 1971. On distance estimators of density in randomly distributed forests. *Biometrics* **27**: 991–1002.